Instituto Tecnológico y de Estudios Superiores de Monterrey Campus Monterrey

**Aplicación de métodos multivariados en ciencia de datos (MA2003B.201)**

# *"E2. Evidencia del Reto"*

Hecho por:

Sergio Alejandro Esparza González – A01625430

Socio Formador:

Sistema de Monitoreo Ambiental (SIMA)

Fecha:

20 de octubre de 2024

# E2. Evidencia del Reto

En este proyecto se realizó un Modelado Predictivo de la Contaminación Atmosférica en Monterrey, así como un Análisis del Impacto Industrial de CFE, Ternium y Cemex. Como se sabe, la disponibilidad de datos confiables de la contaminación que es emitida por las principales industrias de la zona metropolitana puede ser limitada, y para el Sistema de Monitoreo Ambiental (SIMA) es muy importante verificar la exactitud de la información proporcionada por las industrias, así que es fundamental contar con un mecanismo que pueda validar la veracidad y fiabilidad de los datos reportados.

Se tuvo como objetivo determinar los contaminantes predominantes en la atmósfera, considerando su distribución geográfica, y establecer una correlación con la ubicación de las plantas industriales más relevantes dentro de la Zona Metropolitana de Monterrey, puesto que si se puede identificar la contaminación de la industria, se pueden comparar los datos generados con los reportados para poder así verificar su veracidad.

Se eligieron los sectores industriales de: Generación de energía eléctrica (CFE), Metalurgia (Ternium), y Cementeras (Cemex), debido a que según el Programa Integral de Gestión de Calidad del Aire, el sector de la generación de energía eléctrica genera más óxidos de nitrógeno (NOx), monóxido de carbono (CO) y partículas en suspensión atmosférica de 10 micrómetros de diámetro (PM10); el segundo sector de metalurgia genera también mucho NOx, CO y PM10; y el sector de las cementeras genera más NOx, CO y dióxido de azufre (SO2). (Tabla 1).

| Sector | Emisiones de contaminantes año 2018 (t/año) | | | | | | |
|---|---|---|---|---|---|---|---|
| | $PM_{10}$ | $PM_{2.5}$ | $SO_2$ | CO | $NO_x$ | COV | $NH_3$ |
| Fuentes fijas | 59% | 58% | 99% | 4% | 21% | 10% | 3% |
| Generación de energía eléctrica | 10.0% | 14.5% | 0.1% | 1.9% | 8.2% | 0.2% | 0.9% |
| Metalúrgica (incluye la siderúrgica) | 9.4% | 10.4% | 0.2% | 0.5% | 1.9% | 0.4% | 0.6% |
| Petróleo y petroquímica | 8.0% | 6.4% | 93.5% | 0.4% | 2.8% | 0.8% | 0.5% |
| Minerales no metálicos | 7.3% | 6.8% | 0.1% | 0.1% | 0.3% | 0.1% | 0.1% |
| Extracción/Beneficio minerales no metálicos | 7.2% | 4.4% | 0.0% | 0.0% | 0.1% | 0.0% | 0.0% |
| Automotriz | 6.6% | 4.7% | 0.0% | 0.3% | 0.4% | 2.0% | 0.1% |
| Vidrio | 2.4% | 2.4% | 1.0% | 0.2% | 2.7% | 0.2% | 0.2% |
| Química | 2.0% | 1.9% | 0.3% | 0.1% | 1.8% | 2.2% | 0.1% |
| Cemento y cal | 1.9% | 1.7% | 3.8% | 0.6% | 1.8% | 0.0% | 0.1% |
| Accesorios, aparatos eléctricos y equipos de generación eléctrica | 1.1% | 1.3% | 0.0% | 0.0% | 0.1% | 0.5% | 0.0% |
| Metálico | 0.8% | 0.9% | 0.0% | 0.1% | 0.2% | 0.6% | 0.1% |
| Alimentos y bebidas | 0.7% | 0.4% | 0.3% | 0.0% | 0.1% | 0.0% | 0.0% |
| Plástico y hule | 0.5% | 0.4% | 0.0% | 0.0% | 0.0% | 0.6% | 0.0% |
| Celulosa y papel | 0.4% | 0.5% | 0.0% | 0.2% | 0.4% | 0.1% | 0.1% |

*Tabla 1.* Porcentaje de participación por sector en las emisiones totales del inventario del año 2018.

En la exploración de los datos, primero se unieron las bases de datos de los distintos años en una sola, la columna “fecha” se dividió en “date” y “time” para poder facilitar la creación de mapas, y se agregó una columna llamada “location” para poder identificar la estación de origen de cada año. Posteriormente, la columna “location” fue sustituida por las coordenadas de latitud y longitud para poder representar las ubicaciones en los mapas. Los datos atípicos fueron reemplazados por valores nulos para su tratamiento, y después de limpiar los datos, se llevó a cabo un análisis de distribución estandarizada de las variables relevantes.

En cuanto a los datos faltantes, se aplicaron procesos de imputación y normalización para mejorar su calidad, utilizando algoritmo de Imputación Múltiple con Ecuaciones

Encadenadas (MICE) mediante regresión bayesiana para poder completar los valores faltantes, resultando efectivo según los análisis realizados de manera posterior (figuras 1 y 2).

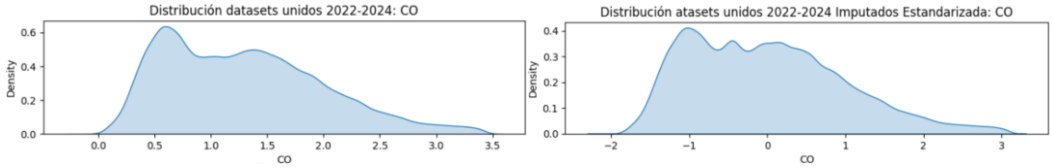


*Figura 1*. Distribución de los datos sin procesar vs. datos imputados y estandarizados.

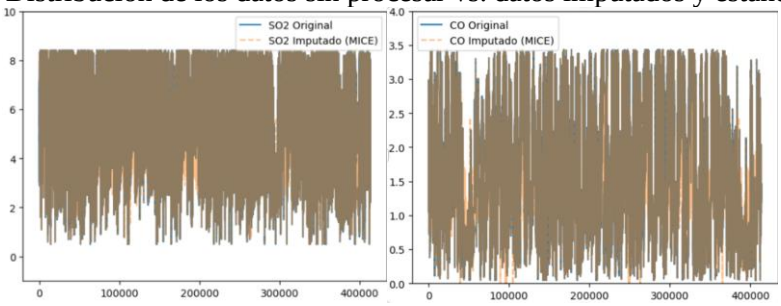


*Figura 2*. Efectividad de los datos imputados para SO2 y CO.

En el análisis de datos, se realizó el algoritmo de K vecinos más cercanos (KNN) para poder agrupar las plantas industriales en función de los contaminantes que las afectan de manera más significativa (figura 3), utilizando CO, NOx, PM10, PM2.5 y SO2.

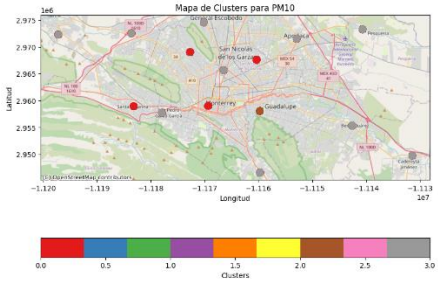


*Figura 3*. Mapa de clústeres para PM10.

Posteriormente, se hizo un análisis de componentes principales (PCA), pues ayuda a reducir la cantidad de dimensiones sin perder información importante, haciendo más fácil analizar los contaminantes al identificar patrones clave. En este análisis, para cada empresa se utilizó una combinación de estaciones de medición diferentes de acuerdo a su cercanía a los puntos generadores de contaminación (tabla 2). Se utilizaron cuatro estaciones para cada empresa.

| Planta | Componente Principal | Contaminantes Dominantes | Variabilidad |
|---|---|---|---|
| CFE | CP1 | PM10, PM2.5, NOx, CO | 40% |
| | CP2 | CO, SO2 | 20% |
| TERNIUM | CP1 | PM10, PM2.5, NOx | 40% |
| | CP2 | CO, SO2 | 20% |
| CEMEX | CP1 | PM10, PM2.5, NOx | 40% |
| | CP2 | CO, SO2 | 20% |

*Tabla 2*. Resultados del PCA.

Finalmente, se realizaron mapas de degradados utilizando interpolación cúbica, que es una técnica que ajusta una función polinómica de tercer grado entre puntos de datos conocidos, pudiendo modelar variaciones suaves y naturales en los datos de contaminación, evitando cambios bruscos no realistas. En este caso, se emplearon los valores conocidos de contaminantes como CO, junto con las condiciones del viento como velocidad para poder inferir la distribución de los niveles de contaminación en áreas intermedias, permitiendo crear una malla de estimaciones que refleje de manera más precisa la calidad del aire en ubicaciones donde no se encuentren sensores (figura 4).

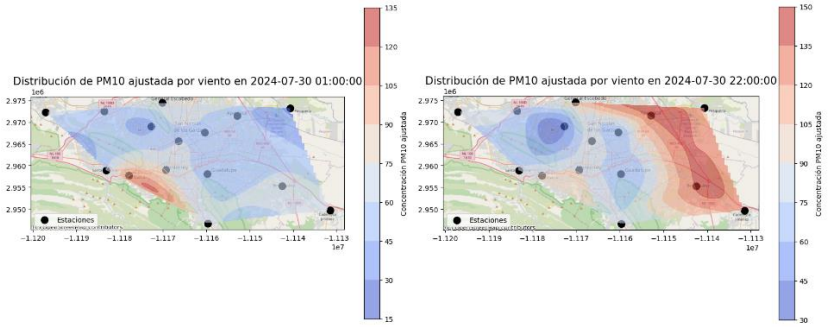


*Figura 4.* Distribución de PM10 en Monterrey, 30 de julio de 2024 de 01:00 a 22:00.

En conclusión, el proyecto logró integrar de manera eficaz bases de datos estructuradas y no estructuradas al combinar registros históricos con coordenadas geográficas para una mejor representación de los datos mediante mapas y análisis visuales. Al utilizar métodos de imputación como MICE, se garantiza la calidad y la completitud de los datos, minimizando el impacto de los valores faltantes. En cuanto al análisis exploratorio y predictivo, se utilizaron algoritmos adecuados para extraer información relevante, como el KNN para agrupar las fuentes industriales por contaminante, y el PCA para optimizar la cantidad de variables relevantes, facilitando la interpretación y reducción dimensional sin pérdida significativa de información. La interpolación cúbica brindó un enfoque avanzado en la estimación de la contaminación en áreas no monitoreadas, mostrando una distribución precisa de los contaminantes.

Desde una perspectiva crítica, los resultados muestran la importancia del monitoreo continuo y verificable para validar la información proporcionada de las industrias, mostrando una correlación geográfica entre las plantas industriales y los contaminantes, para poder facilitar la comparación entre los reportes empresariales y las predicciones, ayudando a identificar inconsistencias en la información. Este proyecto es un claro ejemplo de cómo las matemáticas aplicadas y la ingeniería contribuyen de manera significativa a resolver problemas complejos como la contaminación atmosférica al aplicar técnicas estadísticas y herramientas avanzadas para generar información útil para la toma de decisiones rumbo a una gestión más sostenible del medio ambiente en la Zona Metropolitana de Monterrey.

En este curso de Aplicación de métodos multivariados en ciencia de datos fue posible profundizar en herramientas fundamentales para analizar datos complejos, siendo muy relevante el cómo los métodos se conectan de forma práctica con casos reales, permitiendo descubrir patrones y relaciones ocultas entre los datos. Al trabajar con datos reales e interpretar sus implicaciones, se comprende mejor el impacto de la ciencia de datos en distintos campos del mundo real, como la industria y el medio ambiente, lo cual es motivador.